HANDLING ADVERSARIES IN IMAGE RECOGNITION DEEP NEURAL NETWORKS

Pivithuru Thejan Amarasinghe

(209307X)

Degree of Master of Science

Department of Computer Science and Engineering

Faculty of Engineering

University of Moratuwa
Sri Lanka

March 2022

HANDLING ADVERSARIES IN IMAGE RECOGNITION DEEP NEURAL NETWORKS

Pivithuru Thejan Amarasinghe

(209307X)

Thesis/Dissertation submitted in partial fulfilment of the requirements for the degree Master of Science in Data Science

Department of Computer Science and Engineering

Faculty of Engineering

University of Moratuwa
Sri Lanka

March 2022

Declaration

I declare that the thesis as my own work and it doesn't contain any sort of external material without an acknowledgement. Also, according to my knowledge I haven't used any previously published materials in order to prepare the thesis. All the work carried out to prepare the thesis are individual work and do not copy any existing work without an acknowledgement.

I like to grant permission to the University of Moratuwa to reuse or share my work in any sort of format with external parties. Also, I like to keep my right to use my work in future.

Signature: Date:

I supervised the research conducted by the above candidate.

Signature of the supervisor: Date:

Acknowledgments

I wish to sincerely thank my supervisor, Dr. Charith Chitraranjan for providing me with the necessary guidance throughout the project. He provided me with ideas from the beginning of the project and helped me with finding the research idea. Also, special gratitude should be given to our MSc coordinator for encouraging me to full fill my research under dedicated timelines. Also, I would like to extend my gratitude to all the academic staff of the University of Moratuwa for the great contribution they made for me during study. Also, I am grateful to my family members and all the friends who helped me throughout the project.

Abstract

Deep neural networks play a vital role in image recognition. There are so many mission-critical applications that use deep neural networks for image recognition. With the popularization of deep neural networks, attackers have identified their downsides of them when it comes to image recognition. Some ways can create images that can fool even deep neural networks. These images are commonly known as adversarial images. So attackers use these adversarial images to fool image recognition neural networks to develop a negative picture about using neural networks for image recognition. And even sometimes, attackers use these loopholes to conduct criminal activities as well. Keeping all these aspects in mind the idea of the research is to develop a viable solution that can tackle the main two attack techniques. The research will focus on developing adversarial images using main attacking techniques and developing a defense mechanism for those attacks. The defense technique used in the research is a combination of two techniques called adversarial training and defense distillation. As the outcome of the project accuracy of the proposed solution is measured against a typical deep neural network-based image recognition system using data samples containing adversarial images.

Table of Contents

List of Figures

List of Tables